Introducing Undergraduate Students to Spam Filtering, Internet Traffic Data and WWW Clickstream Data Analysis: Three Activities that Work.

Juana Sanchez
UCLA Department of Statistics
jsanchez@stat.ucla.edu

Introduction

In the Spotlight Session, I illustrate the numerical and graphical analysis for the three activities that I summarize below.

1.- Describing and modeling web browsing behavior. ***Activity appropriate for the Introductory Statistics service course during (a) the descriptive data analysis part of the course, including the correlation/regression part, and (b) sampling theory.***

Here is what your requests for web pages of the Statistics Department web site look like in the web server logs. Your IP address, the time when you clicked, your request, a code saying it is ok, the bytes you are requesting and where you are coming from and operating systems, Internet Service providers and software are recorded for every single click you make.

```
61.149.137.109 - - [01/Jun/2004:00:00:12 -0700] "GET /index.php?vol=2 HTTP/1.1" 200 32896
"http://www.jstatsoft.org/index.php?vol=1" "$
64.68.82.14 - - [01/Jun/2004:00:00:21 -0700] "GET /~cochran HTTP/1.0" 200 2991 "-" "Googlebot/2.1
(+http://www.googlebot.com/bot.html)"
127.0.0.1 - - [01/Jun/2004:00:01:30 -0700] "GET /server-status" 200 17082 "-" "-"
80.232.169.174 - - [01/Jun/2004:00:02:22 -0700] "GET /v06/i06/codes/mingcv.m HTTP/1.1" 200 1806 "-"
"tfqsgmsnnpurmbwmgjdyglyogxdpwe"
212.247.91.99 - - [01/Jun/2004:00:03:10 -0700] "GET / HTTP/1.1" 200 21029
"http://members.aol.com/johnp71/javasta3.html"
"Mozilla/4.0 $
```

How can you learn from this observational time data about the behavior of visitors to this web site? Whatever your questions, the web logs have to be cleaned and processed accordingly, and the visitors' clickstream has to be determined. Once that is done, we can process the data further to make them easily managed by students. Suppose we are interested only in having a data set containing which page each visitor has clicked on, the length or total number of pages (including repeats), the first page clicked and the number of unique pages visited. With this data set your students in the Intro Stats class can find out a lot of things. For example, let's see first a few lines of this data set for the msnbc web site.

User session 1: 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 2 1 Entered page 1, clicked twice p1
User session 2: 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 2 1 1 Entered page 2, clicked once p2
User session 3: 0 5 3 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 3 9 3 Entered page 3, visited 3 pages, 9 clicks.

From this, students can find the distribution of length and depth, the most common gate of entry, the popularity of pages, the correlation between depth and length. The datasets are very large and can also be used to illustrate the CLT and other sampling concepts.

2.- Fitting distributions to Internet Packet Traffic Data. *Activity appropriate for (a) the descriptive and probability modules of the introductory statistics course, lower and upper division, and (b) mathematical statistics class.*

Logs of packet traffic arriving to mail, login or ftp servers have formats similar to those of the web server logs above. Processing the data for our students is necessary. Important questions in this area that are relevant in the introductory courses are: what distributions fit the variables "packet size," "number of packets per unit of time" and "time between arrivals of packets? With voice traffic, such as telephone, the last two were consistently Poisson and Exponential respectively, but that is no longer the case with internet packet traffic.

Students can compare the distributions of the data with known simulated distributions with the same parameters , do q-q plots, goodness of fit tests, mle. But more importantly, they learn to appreciate thick tail distributions, models not usually studied such as power laws, which are very prevalent in internet data analysis. Then they can look at the behavior of traffic over time to try to explain the distributions, and learn to distinguish between outliers and thick tail behavior.

3.- Using Bayes Theorem to predict whether a mail message is spam or not. *Activity appropriate for the introduction to probability part of the Introductory Statistics service course, and the part where we summarize categorical variables or do chi-square tests.*

Give students a training corpus (a set of randomly chosen messages known to be spam or good and ask them to determine the proportion of times a word appears in spam messages, and proportion of times a word appears in non-spam messages (including the headings). That is the empirical P(word|spam) and P(word|not-spam). Chi-square tests can be done to determine if the frequencies differ significantly for spam and good mail. But to filter spam, it is more important to determine what is the probability that a mail message arriving to our mail server is spam.

$$P(Spam \mid message) = \frac{P(message \mid spam)P(spam)}{P(message)}$$

$$P(message) = P(mess \mid spam)P(spam) + P(mess \mid not-spam)P(not-spam)$$

P(message|spam)=P(first word|spam)P(second word|spam).....P(last word|spam)
P(message|not spam)=P(first word|not spam)P(second word|not spam).....P(last word|not spam). Once this is computed, give students a random sample of email messages and ask them to classify them as spam or good. Compute the false positives. Discuss limitations of approach.

References:

(1) Sanchez, J. CS-STATS. http://www.stat.ucla.edu/~jsanchez/oid03/csstats/index.htm and references therein.
(2) Sanchez, J and Ye, S. "Internet Data Analysis for the Undegraduate Statistics Curriculum." Forthcoming JSE
(3) Goodman J, Heckerman D. "Fighting Spam with Statistics." Significance, Vol 1, Issue 2. 2004